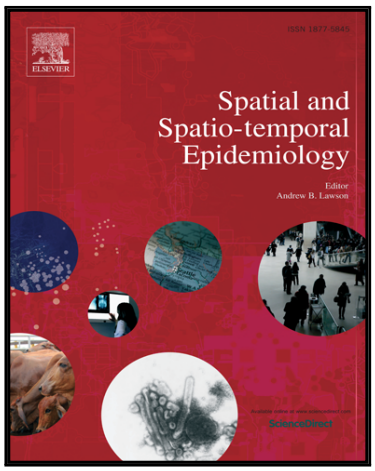

Modelling of spatio-temporal variation in plague incidence in Madagascar from 1980 to 2007

Emanuele Giorgi, Katharina Kreppel, Peter J. Diggle,
Cyril Caminade, Maherisoa Ratsitorahina, Minoarisoa Rajerison,
Matthew Baylis

PII: S1877-5845(16)30030-2
DOI: 10.1016/j.sste.2016.10.001
Reference: SSTE 202

To appear in: *Spatial and Spatio-temporal Epidemiology*

Received date: 16 May 2016
Revised date: 12 September 2016
Accepted date: 12 October 2016

Please cite this article as: Emanuele Giorgi, Katharina Kreppel, Peter J. Diggle, Cyril Caminade, Maherisoa Ratsitorahina, Minoarisoa Rajerison, Matthew Baylis, Modelling of spatio-temporal variation in plague incidence in Madagascar from 1980 to 2007, *Spatial and Spatio-temporal Epidemiology* (2016), doi: 10.1016/j.sste.2016.10.001

Modelling of spatio-temporal variation in plague incidence in Madagascar from 1980 to 2007

Emanuele Giorgi[1], Katharina Kreppel[2], Peter J. Diggle[1], Cyril Caminade[3], Maherisoa Ratsitorahina[4], Minoarisoa Rajerison[4], Matthew Baylis[3,5]

[1] *Lancaster Medical School, Lancaster University, Lancaster, UK*
[2] *Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow, UK*
[3] *Institute of Infection and Global Health, Department of Epidemiology and Population Health, University of Liverpool, Leahurst Campus, Neston, Cheshire, UK*
[4] *Unité Peste, Institut Pasteur de Madagascar, Antananarivo, Madagascar*
[5] *Health Protection Research Unit for Emerging and Zoonotic Infections, University of Liverpool, UK*

October 18, 2016

Abstract

Plague is an infectious disease caused by the bacterium *Yersinia pestis*, which, during the fourteenth century, caused the deaths of an estimated 75 to 200 million people in Europe. Plague epidemics still occur in Africa, Asia and South America. Madagascar is today one of the most endemic countries, reporting nearly one third of the human cases worldwide from 2004 to 2009. The persistence of plague in Madagascar is associated with environmental and climatic conditions. In this paper we present a case study of the spatio-temporal analysis of plague incidence in Madagascar from 1980 to 2007. We study the relationship of plague with temperature and precipitation anomalies, and with elevation. A joint spatio-temporal analysis of the data proves to be computationally intractable. We therefore develop a spatio-temporal log-Gaussian Cox process model, but then carry out marginal temporal and spatial analyses. We also introduce a spatially discrete approximation for Gaussian processes, whose parameters retain a spatially continuous interpretation. We find evidence of a cumulative effect, over time, of temperature anomalies on plague incidence, and of a very high relative risk of plague occurrence for locations above 800 meters in elevation. Our approach provides a useful modelling framework to assess the relationship between exposures and plague risk, irrespective of the spatial resolution at which the latter has been recorded.

Keywords: Cox process; distributed lag-model; Gaussian process; plague; spatio-temporal point pattern.

1 Background

Plague, also known as the "Black Death" during the fourteenth century, is a vector-borne disease caused by the zoonotic bacterium *Yersinia pestis*, usually found in small mammals, especially rodents. It can be transmitted to humans through the bite of an infected flea or, in some cases, through direct contact, inhalation or digestion of infected material. Three different forms of plague can be distinguished: bubonic plague, which is the most common form and characterized by swollen lymph nodes, called "buboes"; septicaemic plague, when the infection spreads from the buboes through the bloodstream; pneumonic plague, the most virulent form, that occurs when the infection, usually from advanced bubonic plague, spreads to the lungs. If diagnosed in time, plague can be effectively treated with antibiotics. If untreated, it can be rapidly fatal.

Today, plague is mainly present in South America, Asia and Africa. Madagascar is one of the most endemic countries, accounting for almost one third of the human cases worldwide from 2004 to 2009 (WHO, 2009). Plague reached the coast of Madagascar in 1898 with steamboats from India (Brygoo, 1966), following the third pandemic that originated in China in 1855. It then spread to the Central Highlands in 1921 where it became endemic. The main host for plague on the island is the black rat, with its fleas acting as the vector. The persistence of plague foci in Madagascar is associated with climatic, environmental and sociological factors (Ben Ari et al., 2011) all of which affect the distribution of flea and rodent populations, and their contact with humans. Among the climatic factors, temperature, rainfall and relative humidity are considered to affect the disease dynamics, mostly through their effects on the survival and reproduction of host and vector (Stenseth et al., 2006). In Madagascar, plague is mainly found in rural areas that provide the preferred habitat of the main host for *Y. pestis*, the black rat *Rattus rattus*. Traditional burial customs and lack of access to affordable health care also impact on the spread and persistence of the disease, but data on these factors is currently lacking. Understanding how all these factors interact and jointly affect human plague incidence requires the use of appropriate multivariate statistical methods.

In this paper we analyse monthly data on district-level plague incidence in Madagascar from January 1980 to December 2007. Motivated by Diggle et al. (2013), we model monthly district-level plague incidence as an aggregated outcome of a spatio-temporally continuous process. The objective of our analysis is to quantify the effects of temperature and precipitation anomalies (henceforth TAs and PAs), and of elevation, on plague incidence whilst accounting for unexplained spatio-temporally structured variation. One of the main features of the analysed data is that temporal variation in incidence dominates its spatial variation, making the joint spatio-temporal analysis of the data computationally cumbersome. To address this issue, we carry out marginal temporal and spatial analyses based on a spatio-temporal log-Gaussian Cox process model. For the spatial analysis, we also introduce a spatially discrete approximation to the underlying continuous Gaussian process, for which inference is computationally more feasible, but whose parameters retain a spatially continuous interpretation.

Kreppel et al. (2014) have previously analysed country-level plague incidence anomalies from 1960 to 2007. They used wavelet analysis to study the relationship between plague and climatic anomalies caused by the El Niño Southern Oscillation (ENSO) and the Indian Ocean Dipole

(IOD). They found evidence of a temporally varying relationship between plague incidence anomalies and TAs, PAs, ENSO and IOD. However, one of the main limitations of their approach is that only marginal associations between plague and climatic risk factors were investigated. Additionally, in their analysis plague incidence was treated as a continuous outcome, which is questionable in view of the low case-counts that are observed throughout the time series. In the review paper by Andrianaivoarimanana et al. (2013), maps of empirical plague incidence show its restriction mainly to the Central Highlands districts. Based on summary descriptive statistics, the authors also mention a threshold of 800 meters in elevation as critical in determining the geographical limits of plague foci. The modelling approach we propose in this paper overcomes the inherent limits of the analysis by Kreppel et al. (2014), and also allows us to validate the statement by Andrianaivoarimanana et al. (2013) on the effect of elevation, in a statistically rigorous fashion.

The structure of the paper is as follows. In Section 2, we briefly describe the different data sources for plague incidence, elevation, TAs, PAs and population density. In Section 3, we first carry out an exploratory analysis to highlight the main features of the data, then develop a spatio-temporal log-Gaussian Cox process for plague incidence and fit the model using Bayesian inference. In Section 4, we report the results of our analysis and check the validity of the modelling assumptions. Section 5 is a concluding discussion.

2 Data

We use the following data sources, including climatic, environmental and demographic datasets, to obtain the relevant explanatory variables and offset quantities.

- *Plague cases.* We use data on all cases of human plague confirmed by a bacteriological test from January 1980 to December 2007. These data were made available by the World Health Organisation Plague Reference Laboratory of the Institut Pasteur de Madagascar.
- *Population density.* Spatial population density estimates for Madagascar for the year 2010 were derived from the GPWv4 dataset, released by the Center for International Earth Science Information Network, Columbia University (Doxsey-Whitfield et al., 2015). We use 5-yearly country-level human population growth estimates from the United Nations to estimate the total population of Madagascar from 1980 to 2007, based on the latest population census conducted in 1993.
- *Temperature and precipitation.* Monthly average values of temperature and precipitation were obtained from the NCEP-NCAR reanalysis (Kalnay et al., 1996). These were available as averaged values over a large domain including Madagascar [42.18° E−49.68°E; 24.76°S−11.42°S].
- *Elevation.* Spatial estimates of elevation were obtained from the ETOP01 dataset made available by the National Oceanic and Atmospheric Administration (Amante & Eakins, 2009).

3 Model formulation

3.1 Exploratory analysis

In Madagascar, plague is mainly found on the Central Highlands but, as with all infectious diseases, it can also travel to non-endemic areas with its host. Cases of disease may be recorded, therefore, outside of plague's natural (endemic) range. In Figure 1, the average annual incidence of plague, from 1980 to 2007, is shown for each district, where the purple line encompasses regions with elevation 800 meters or more. At least one case was reported in only 44 of the 110 districts, of which only two lie completely below 800 meters. These districts are Mahajanga Rural and Mahajanga Urban on the west coast with a total of 410 cases occurring between August 1991 and November 1999.

Figure 2 shows the boxplots of Madagascar-wide plague incidence by month (left panel) and by year (right panel). The presence of a strong seasonal component is evident, with most plague cases occurring during the warm and wet austral summer season between October and March. An increasing trend in incidence, with an associated increase in variation, is observed until around 1997 after which incidence decreased.

We generate TAs and PAs by subtracting from each datum the mean for the month in question over the years, from 1979 to 2007. The resulting TA and PA time series are shown in Figure 3.

3.2 Spatio-temporal log-Gaussian Cox process

A spatio-temporal Cox process (Cox, 1955) is defined by the two following postulates.

P1 Let $\tilde{\Lambda} = \{\tilde{\Lambda}(x,t), x \in \mathbb{R}^2, t \in \mathbb{R}\}$ denote a non-negative valued stochastic process.

P2 Conditionally on $\tilde{\Lambda}(x,t), x \in \mathbb{R}^2, t \in \mathbb{R}$, the point process is an inhomogenous Poisson process with intensity $\tilde{\Lambda}(x,t)$.

Following the modelling framework developed by Diggle et al. (2013), we then model the district-level plague counts as an aggregated outcome of the spatio-temporal process $\tilde{\Lambda}$. More specifically, let Y_{it} denote the number of plague cases in the i-th district at time t, for $i = 1, \ldots, d = 110$ and $t = 1, \ldots, T = 306$, where $t = 1$ indicates January 1980. We assume that, conditionally on $\tilde{\Lambda}(x,t)$, the Y_{it} are mutually independent Poisson variables with means

$$\begin{aligned} \mu_{it} &= \int_{\mathcal{R}_i} \int_{t-1}^{t} \tilde{\Lambda}(u,v)\, dudv \\ &= \int_{\mathcal{R}_i} \int_{t-1}^{t} \lambda_0 n(u,v) \Lambda(u,v)\, dudv, \end{aligned} \tag{1}$$

where $\mathcal{R}_i$ is the area encompassed by the boundaries of the i-th district, $n(x,t)$ is an offset representing the number of susceptibles and λ_0 is an unknown positive constant. The intensity

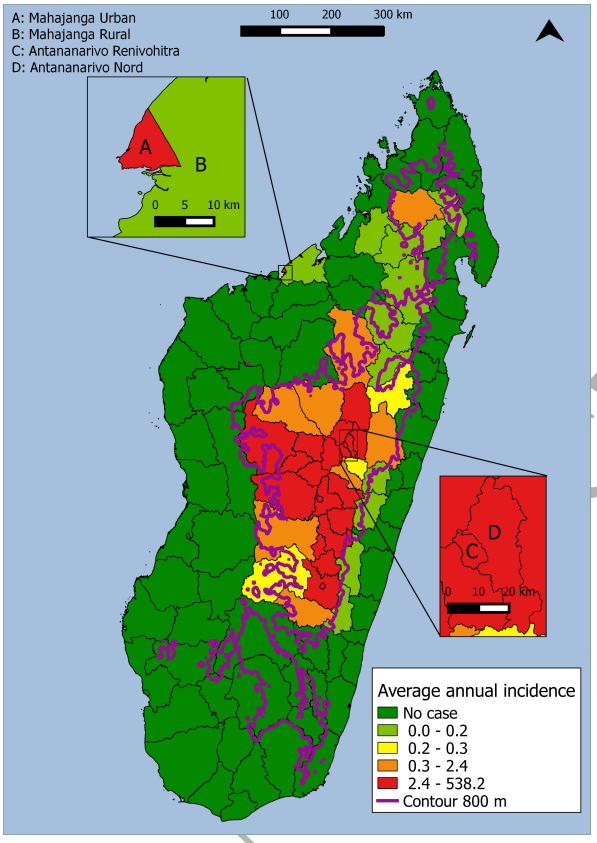

Figure 1: Map of the observed average annual incidence per 100,000 people in each district. The purple line corresponds to the contour of 800 meters in elevation.

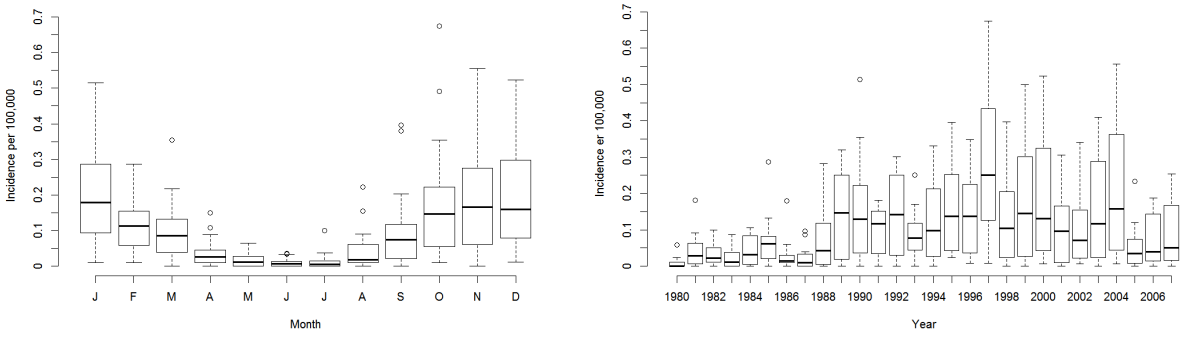

Figure 2: Boxplots of average human plague incidence in Madagascar per 100,000 by month (left panel) and year (right panel).

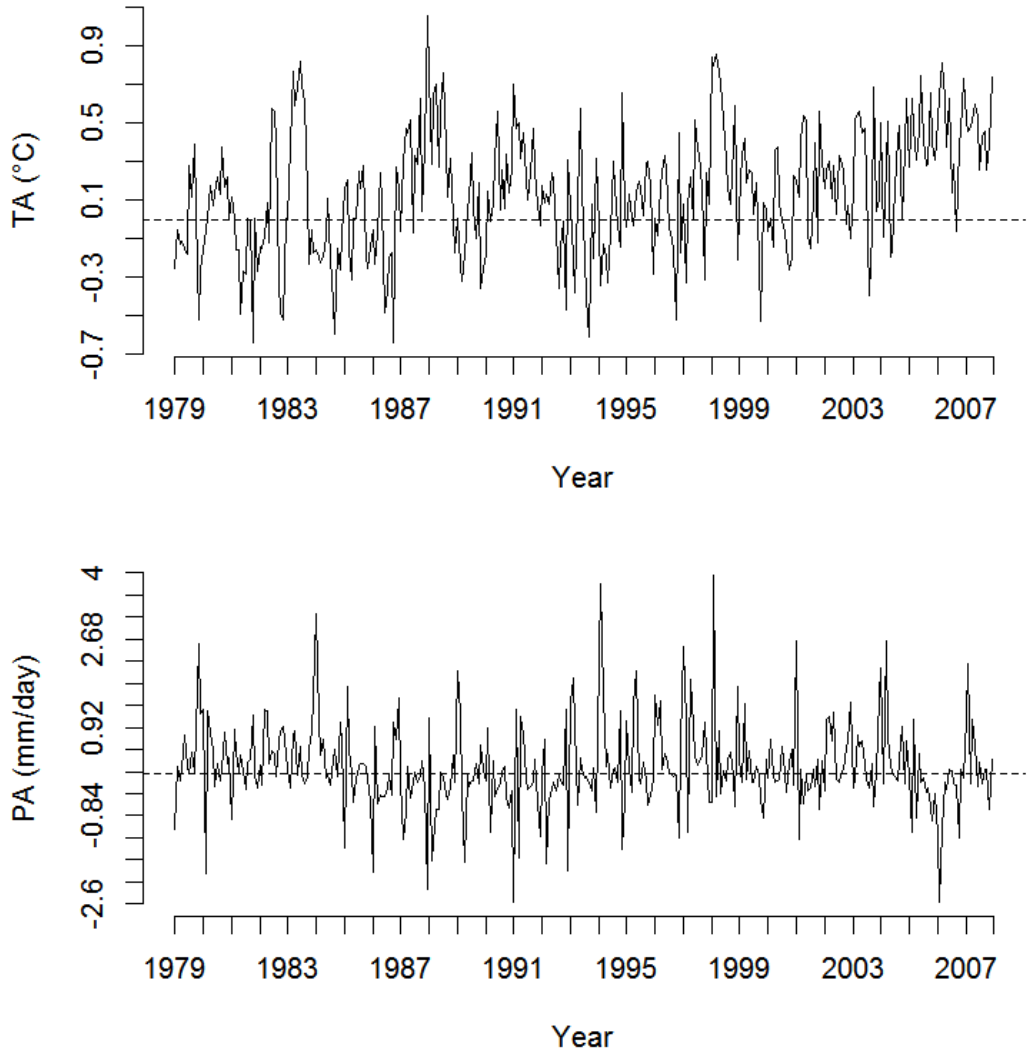

Figure 3: Monthly temperature (upper panel) and precipitation (lower panel) anomalies from January 1979 to December 2007. The dashed line corresponds to a zero anomaly value.

of the inhomogenous Poisson process, given by $\tilde{\Lambda}(x,t) = \lambda_0 n(x,t)\Lambda(x,t)$, is defined on $\mathcal{M} \times [0,T]$, with $\mathcal{M} = \bigcup_{i=1}^{d} \mathcal{R}_i$. We calculated the number of susceptibles as

$$n(x,t) = \frac{N(t)p(x)}{\int_{\mathcal{M}} p(u)\, du}$$

where $p(x)$ is the population density at location x and the $N(t)$ are five-yearly population estimates for the country of Madagascar at time t. Annual district-level population counts were not available; see Section 2 for more details on the data sources.

We then assume that

$$\Lambda(x,t) = a(x)b(t)V(x,t) \tag{2}$$

where $V(x,t)$ is a stationary and isotropic log-Gaussian process, $a(x)$ and $b(t)$ are deterministic functions specified as regressions on observed spatial and temporal covariates, respectively. The spatial regression on elevation is defined as follows

$$\log\{a(x)\} \;=\; \begin{cases} \gamma & \text{if elevation at } x \text{ is greater than 800 meters,} \\ 0 & \text{otherwise} \end{cases}.$$

We specify three different regression models for $b(t)$ as follows.

- *Model 1.*

$$\begin{aligned} \log\{b(t)\} \;=\; & \sum_{l=0}^{12} a_l \mathrm{PA}(t-l) + \sum_{l=0}^{12} b_l \mathrm{TA}(t-l) + \\ & \beta_1 t + \beta_2 t^2 + \beta_3 \cos\{2\pi t/12\} + \beta_4 \sin\{2\pi t/12\}. \end{aligned}$$

- *Model 2.* As Model 1 but with no effect of PAs, i.e. $a_l = 0$ for $l = 0, \ldots, 12$.
- *Model 3.* As Model 2 but with $b_l = b_0 \exp(-l/r), r > 0$ for $l = 0, \ldots, 12$.

In Model 1, PAs and TAs modulate, through the lagged effects a_l and b_l, the yearly cycle of plague incidence as defined by the linear combination of sine and cosine functions. Model 2 is nested within Model 1 and assumes no effect of PAs on plague incidence. Finally, Model 3 constraints the lagged coefficients b_l, in Model 2, to follow an exponentially decaying function at rate $1/r$. The removal of PAs in Model 2 and Model 3 allows us to assess the importance of PAs effects on plague incidence. This is motivated by the findings of Kreppel et al. (2014) who found a higher degree of association between plague and temperature.

The spatio-temporal stochastic component is modelled as $V(x,t) = \exp\{S(x) + U(t)\}$ where $S(x)$ and $U(t)$ are stationary zero-mean Gaussian processes. Since the main goal of our analysis is to make inference on the regression parameters of $a(x)$ and $b(t)$ while accounting for unmeasured spatio-temporal risk factors, we conduct separate marginal temporal and

spatial analyses as follows. Let $\tilde{p}(x) = p(x)/\int_{\mathcal{M}} p(u)\,du$, then the marginal spatial intensity function of the Poisson process is given by

$$\begin{aligned}\int_0^T \lambda_0 n(x,v)\Lambda(x,v)\,dv &= \lambda_0 \tilde{p}(x)a(x)\exp\{S(x)\}\int_0^T N(v)b(v)\exp\{U(v)\}\,dv \\ &= \alpha_1 \tilde{p}(x)a(x)\exp\{S(x)\}, \end{aligned} \tag{3}$$

where $\alpha_1 = \lambda_0 \int_0^T N(v)b(v)\exp\{U(v)\}\,dv$. Similarly the marginal temporal intensity is

$$\begin{aligned}\int_{\mathcal{M}} \lambda_0 n(u,t)\Lambda(u,t)\,du &= \lambda_0 N(t)b(t)\exp\{U(t)\}\int_{\mathcal{M}} \tilde{p}(u)a(u)\exp\{S(u)\}\,du \\ &= \alpha_2 N(t)b(t)\exp\{U(t)\}, \end{aligned} \tag{4}$$

where $\alpha_2 = \lambda_0 \int_{\mathcal{M}} \tilde{p}(u)a(u)\exp\{S(u)\}\,du$. We assume exponential covariance functions

$$\mathrm{cov}\{S(x), S(x')\} = \sigma^2 \exp\{-\|x - x'\|/\phi\},$$

where $\|\cdot\|$ denotes the Euclidean distance, and

$$\mathrm{cov}\{U(t), U(t')\} = \nu^2 \exp\{-|t - t'|/\psi\}.$$

As a consequence of this approach, we do not need to fully specify the joint distribution of $S(x)$ and $U(t)$, thus we do not make any assumption about their dependence or otherwise. This implies that, whilst $V(x,t)$ is defined as a separable process in $S(x)$ and $U(t)$, its correlation function is not necessarily separable. However, we do not pursue prediction of the process $V(x,t)$, in which case additional assumptions have to be made about the joint distribution of $S(x)$ and $U(t)$.

3.3 Inference

We carry out parameter estimation using Bayesian inference. Prior specifications are given in the Appendix.

For the temporal analysis, we developed a Markov Chain Monte Carlo (MCMC) algorithm in which covariance parameters, regression coefficients and random effects are updated separately using a random walk Metropolis-Hasting algorithm, a Gibbs sampler and a Hamiltonian Monte Carlo procedure, respectively.

For the spatial analysis, inference on the continuous process $S(x)$ is computationally unwieldy, especially within districts where no case of human plague has ever been reported. For this reason we approximate $S(x)$ with its district-wide average, i.e.

$$S(x) \approx S_i = \frac{1}{|\mathcal{R}_i|}\int_{\mathcal{R}_i} S(u)\,du, x \in \mathcal{R}_i, \tag{5}$$

where $|\mathcal{R}_i|$ is the area of $\mathcal{R}_i$. Since the integral of a Gaussian process is still a Gaussian process, the joint distribution of $(S_1, \ldots, S_d)$ is multivariate Gaussian with mean zero and covariance matrix with (h,k)-th entry

$$\frac{\sigma^2}{|\mathcal{R}_h||\mathcal{R}_k|}\int_{\mathcal{R}_h}\int_{\mathcal{R}_k} \exp\left(-\frac{\|u - u'\|}{\phi}\right)\,dudu'.$$

We approximate the above integral using a quadrature procedure based on a 7 by 7 km regular grid covering the whole of Madagascar. We then developed an MCMC algorithm in which (σ^2, ϕ) and $(\log \alpha_1, \gamma, S_1, \ldots, S_d)$ are updated separately. The posterior of (σ^2, ϕ) is approximated over a discrete set of points that are identified using an INLA procedure (Rue, Martino & Chopin, 2009). The main reason for this is that it allows us to pre-compute all the required quantities for a significantly faster update of our MCMC algorithm and, also, propose new values for (σ^2, ϕ) in regions of high posterior density. We use a random walk Metropolis-Hastings algorithm for $(\log \alpha_1, \gamma, S_1, \ldots, S_d)$ by proposing a new value, at each iteration, from a multivariate Gaussian with covariance matrix given by the inverse of the negative Hessian computed at the mode, for given (σ^2, ϕ).

Details of the MCMC algorithms are given in the Appendix.

4 Results

4.1 Temporal analysis

For each of the three models, we simulated 35,000 samples and retained every 6-th sample after a burn-in of 5,000. Trace-plots and other diagnostic checks show good mixing of the MCMC with nearly independent samples for the regression parameters and temporal random effects. Table 1 shows posterior summaries for the three models.

Table 1: Posterior mean, standard deviation (SD) and 95% credible intervals (CI) for each of the parameters in Model 1, 2 and 3, excluding PAs and TAs effects. The deviance information criterion (DIC) is also reported for each of the three models.

	Model 1			Model 2			Model 3		
Term	Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI
β_1	0.023	0.004	(0.015, 0.032)	0.023	0.004	(0.014, 0.031)	0.023	0.004	(0.015, 0.031)
$\beta_2 \times 10^4$	-5.306	1.199	(-7.731, -2.901)	-5.033	1.180	(-7.376, -2.687)	-5.087	1.125	(-7.327, -2.851)
β_3	0.108	0.099	(-0.083, 0.305)	0.098	0.092	(-0.082, 0.279)	0.105	0.080	(-0.050, 0.263)
β_4	1.367	0.105	(1.169, 1.576)	1.351	0.092	(1.175, 1.535)	1.408	0.086	(1.241, 1.577)
ν^2	0.496	0.090	(0.348, 0.697)	0.492	0.093	(0.342, 0.704)	0.472	0.087	(0.336, 0.673)
ψ	2.913	0.693	(1.840, 4.508)	2.974	0.724	(1.899, 4.732)	2.904	0.713	(1.879, 4.528)
r	-	-	-	-	-	-	2.482	2.199	(0.211, 7.108)
DIC	139404.6			139402.2			139403.9		

Figure 4 shows the plots of the posterior means for a_l and b_l against l. In the case of PAs, the a_l coefficients appear to follow an erratic pattern around zero, suggesting a weak association between plague incidence and PAs. In the case of TAs, the unconstrained b_l, from Model 2, follow an approximately increasing pattern that seems to level off for $l \geq 3$. The difference in DIC amongst the models is modest, with Model 2 having the lowest value in DIC. However, in the right panel of Figure 4, we do not observe strong evidence against the exponential fit to the unconstrained b_l, and we adopt this as our preferred model. Significant departures from the exponential curve are only observed at lags 1 and 10.

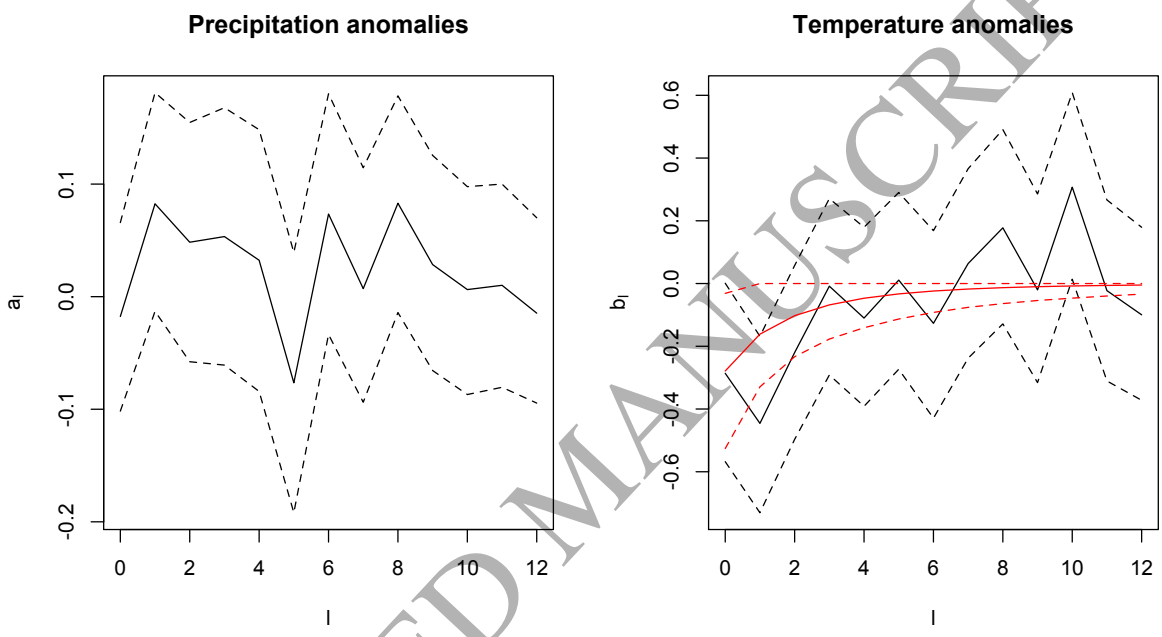

Figure 4: Plots of the posterior mean of a_l (left panel) and b_l (right panel) against l, for each of the fitted models. In the right panel the black solid line corresponds to the posterior mean of the unconstrained b_l from Model 2, whilst the red solid line corresponds to the posterior mean of $b_l = b_0 \exp\{-l/r\}$ from Model 3. The dashed lines in each panel are 95% credible intervals.

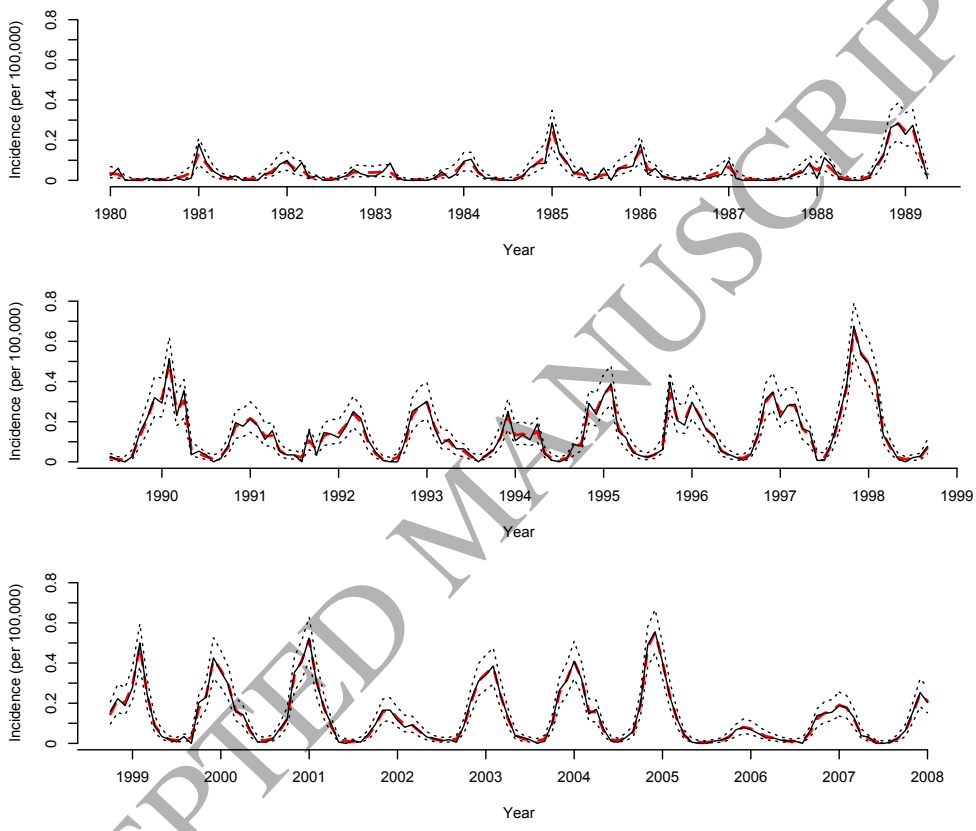

Figure 5: Plots of the observed (solid line) and predicted (dashed red line) incidence per 100,000 people from Model 3, with 95% predictive intervals (dotted lines).

Figure 5 shows the predicted incidence from Model 3. The corresponding autocorrelogram for the residuals, computed as the difference between predicted and observed incidence, shows no residual temporal correlation.

4.2 Spatial analysis

We iterated the MCMC algorithm 1,100,000 times, retaining every 100-th sample after a burn-in of 100,000 samples. The resulting 10,000 posterior samples show good mixing, with almost coincident empirical cumulative density functions for the first and second 5000 samples for each of the model components.

Table 2: Posterior mean, standard deviation (SD) and 95% credible intervals (CI) for the model parameters of the spatial analysis.

Term	Mean	SD	95% CI
γ	6.594	1.515	(4.073, 10.068)
σ^2	21.748	11.465	(10.084, 53.493)
ϕ	134.192	85.695	(50.083, 357.317)

Table 2 shows posterior summaries of the model parameters. We estimate that locations above 800 meters in altitude have a risk of plague occurrence about $e^{6.594} \approx 731$ times higher than those below. The incidence of vector-borne diseases often declines with increasing elevation, as cooler conditions tend to be less favourable for the vectors themselves, and also extend the time required for development of the disease-causing pathogen within them. Plague in Madagascar is unusual, therefore, in that it is most prevalent in the cool highland areas of the country, and is largely absent (with the exception of the coastal town on Mahajanga) in the warm, low-lying and coastal regions (Andrianaivoarimanana et al., 2013). The explanation may lie in the interactions of the two flea vector species in Madagascar. The endemic flea species, *Synopsyllus fonquerniei*, is considered to be a better plague vector than the exotic species, *Xenopsylla cheopis*. The exotic *X. cheopis* has been shown, under laboratory conditions, to survive better at warmer conditions than *S. fonquerniei* (Kreppel et al., 2016). It is possible, therefore, that plague is transmitted most readily in Madagascar only in the cool, highland areas which are climatically suitable for the endemic *S. fonquerniei*.

The estimated range of the spatial correlation, defined as the distance at which the spatial correlation function takes the value 0.05, is about 401 km. Figure 6 shows the predicted number of cases, obtained from the posterior samples, against the observed number of cases for each district. There is a very high concordance between predicted and observed counts, except for the counts reported in Antananarivo-North and Antananarivo Renivohitra. These two contiguous districts are the major conurbation in the country where plague risk is likely to be affected by socio-economic factors, which are not necessarily spatially structured. Both districts are within the capital city boundaries where living conditions range from affluent areas to poverty-stricken areas with cramped living conditions.

Figure 7 shows the posterior mean of the spatial residuals S_i and the predictive probabilities

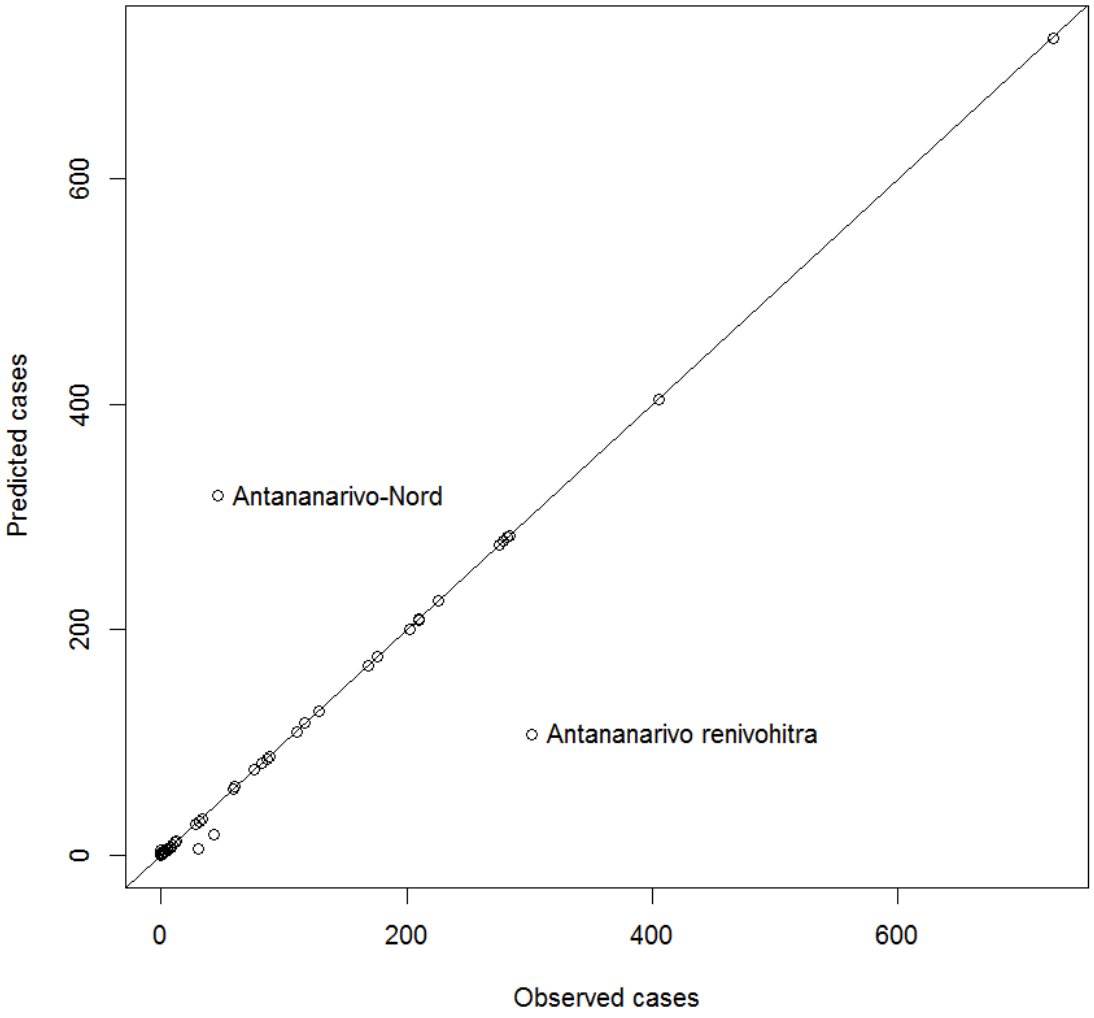

Figure 6: Scatter plot of the predicted against observed number of plague cases in each district. The predicted cases are obtained from the posterior distribution of the model. The solid line corresponds to the identity line.

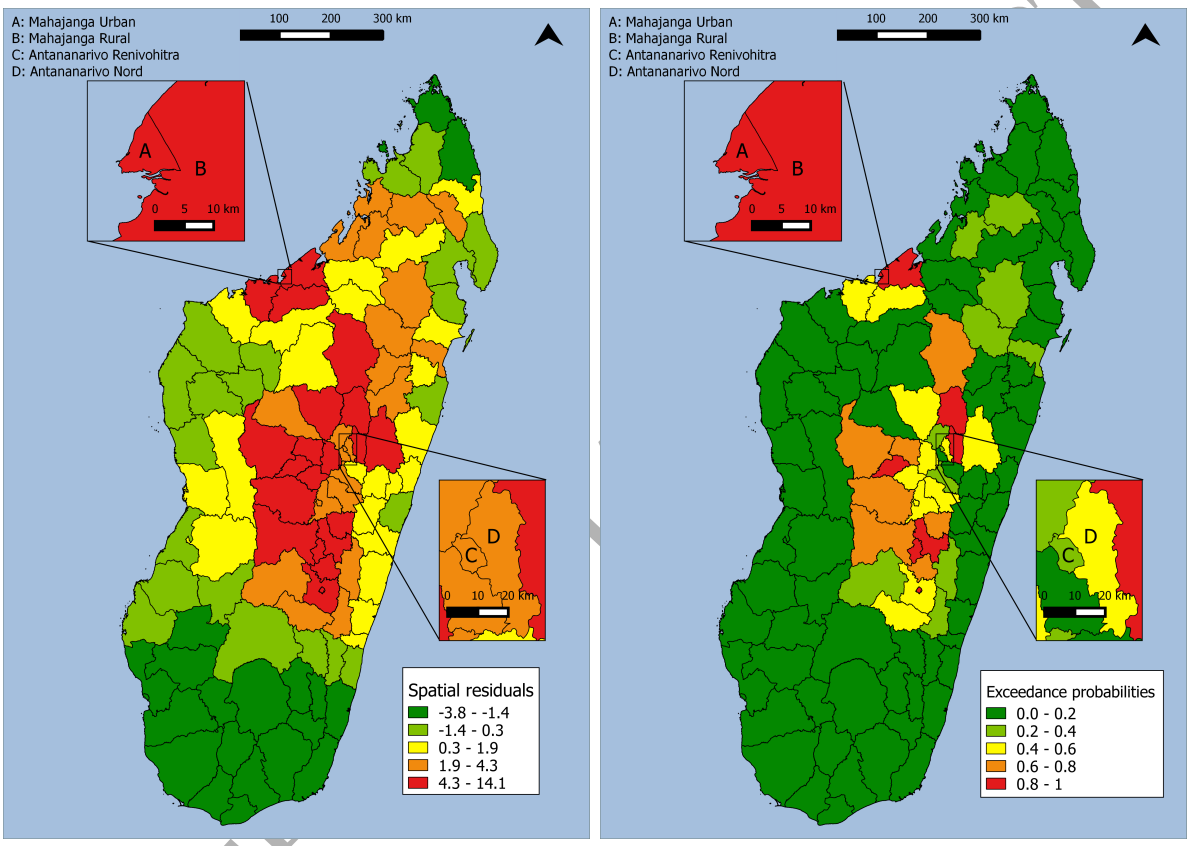

Figure 7: Maps of the posterior mean of the spatial residuals S_i (left panel) and the predictive probabilities of S_i exceeding 4 (right panel).

of S_i exceeding 4, for all i. Districts coloured in red are highly likely associated with a relative risk no less than $e^4 \approx 54.60$. Among these districts, we also find Mahajanga Rural and Mahajanga Urban on the west coast, which lie well below 800 meters. However, due to the higher population density in Mahajanga Urban than in Mahajanga Rural, about 404 plague cases are predicted in the former and only 5 in latter.

4.3 Model diagnostics

In Section 3.2, we made the assumption that $V(x,t) = \exp\{S(x)\}\exp\{U(t)\}$, i.e. that $V(x,t)$ is separable into the processes $S(x)$ and $U(t)$. If this assumption does not hold, the main consequence, in the context of our analysis, would be that the purely spatial and temporal correlation functions cannot be freely defined. More specifically, if $V(x,t) \neq \exp\{S(x)\}\exp\{U(t)\}$, equations (3) and (4) are then replaced by

$$\begin{aligned}\int_0^T \lambda_0 n(x,v)\Lambda(x,v)\,dv &= \lambda_0 \tilde{p}(x)a(x)\int_0^T N(v)b(v)V(x,v)\,dv \\ &= \lambda_0 \tilde{p}(x)a(x)V_1(x) \qquad (6)\end{aligned}$$

and

$$\begin{aligned}\int_{\mathcal{M}} \lambda_0 n(u,t)\Lambda(u,t)\,du &= \lambda_0 N(t)b(t)\int_{\mathcal{M}} \tilde{p}(u)a(u)V(u,t)\,du \\ &= \lambda_0 N(t)b(t)V_2(t). \qquad (7)\end{aligned}$$

In this case, the specification of a stochastic model for $V_1(x)$ would then impose constraints on the set of valid correlation functions that can be used for $V_2(t)$, and vice-versa. More importantly, assuming exponential correlation functions for both $\log\{V_1(x)\}$ and $\log\{V_2(t)\}$ would be an invalid modelling choice, unless $V(x,t)$ is separable into a purely spatial and purely temporal processes.

We then test the validity of the adopted correlation functions as follows. We use $\hat{\eta}$ and $\hat{s}$ to denote the mean and standard deviation of the predicted counts, based on the explanatory variables only. To simplify the notation, we omit any reference to time and space, as we apply the same procedure in both cases. More specifically, in the temporal analysis $\hat{\eta}$ corresponds to the point-wise posterior estimate of $\alpha_2 \int_{t-1}^{t} N(u)b(u)\,du$, for a given time t, with $b(\cdot)$ as defined in Model 3 of Section 3, whilst in the spatial analysis $\hat{\eta}$ estimates $\alpha_1 \int_{\mathcal{R}} \tilde{p}(v)a(v)\,dv$, for a given district $\mathcal{R}$. We then compute the residuals,

$$e = \frac{y - \hat{\eta}}{\hat{s}}, \qquad (8)$$

where y is an observed count (for a single month, in the temporal analysis, and for a single district in the spatial analysis).

We then carry out simulations under the fitted models as follows. For each posterior sample, we first simulate the structured random effects, say E, using the fitted covariance function, then simulate counts, say $\tilde{y}$, from a Poisson distribution with mean $\hat{\eta}\exp\{E\}$. For each $\tilde{y}$, we

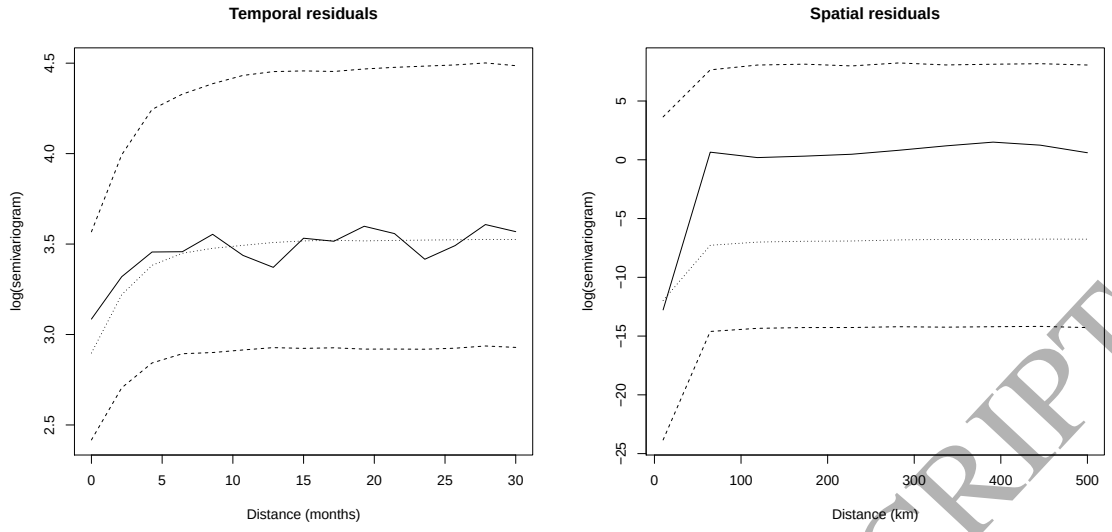

Figure 8: Semi-variograms based on the observed (solid line) and simulated residuals (dotted line) under the assumed temporal (left panel) and spatial (right panel) structure of the model. In each panel, the dashed lines corresponds to the 95% credible intervals obtained from the semi-variograms of the simulated residuals. To facilitate the display of the semi-variograms, these are plotted on the log-scale.

obtain the simulated residuals $\tilde{e} = (\tilde{y} - \hat{\eta})/\hat{s}$ and compute the semi-variogram based on these. For the spatial analysis, distances are computed using the centroids of the districts.

Figure 8 shows the 95% credible intervals (dashed lines) and the median (dotted lines) semi-variograms based on $\tilde{e}$, and the observed (solid lines) semi-variogram. The results are compatible with the assumption of exponential correlation functions for both the temporal and spatial stochastic components. We also note that in the spatial analysis the uncertainty in the simulated semi-variograms is very high, as shown by the very wide credible intervals.

5 Discussion

We developed a log-Gaussian Cox process model in order to quantify the association of plague incidence with PAs, TAs and elevation. Since fitting the joint spatio-temporal model was computationally intractable, we then carried out marginal temporal and spatial analyses. Using distributed-lag models with unconstrained and constrained coefficients, we found evidence of a cumulative effect of temperature anomalies. Also, the results support the findings of Andrianaivoarimanana et al. (2013), denoting plague endemic regions as largely restricted to areas with elevation over 800 meters.

Our approach provides a useful modelling technique for applications where the scientific focus

is more on the relationship of disease incidence with spatial and temporal risk factors, than on testing mechanistic hypotheses. In particular, it also allows us to model both exposures and disease risk as spatially continuous processes, irrespective of the resolution at which incidence data have been recorded. This has the main benefit that assessment of the regression relationships is not affected by the bias induced by the ecological fallacy (Wakefield & Lyons, 2010), in contract to other approaches based on Markov structures (e.g. Held et al. (2005)). Also, in our approach the correlation structure between districts is derived from a spatially continuous stochastic process. In our view, this is a more realistic assumption than a Markov field structure, whereby specification of the spatial model, and therefore the interpretation of regression parameters, is tied to a specific, and arguably arbitrary, partition of the study-region into discrete spatial units.

Our analysis has some limitations due to the lack of detailed climatic and demographic data for the time period considered. Climatic anomalies could only be used as risk factors if uniformly applied to the whole of Madagascar. The susceptible population is estimated by allowing the total population to change only every five years (due to the frequency of the demographic census) and assuming a constant spatial distribution of relative population density. Finally, all the explanatory variables used in the model are the output of either mathematical models or spatial smoothing techniques.

We modelled the residual stochastic component of the model as a separable process, $V(x,t) = \exp\{S(x)\}\exp\{U(t)\}$, with each of $S(x)$ and $U(t)$ having an exponentially decaying correlation function. The diagnostic checks reported in Section 4.3 did not find evidence against this assumption, although they were restricted to checking the marginal spatial and temporal properties of $V(x,t)$. We made the separability of the residual spatio-temporal process assumption for pragmatic reasons, to avoid the computational complexity that would have been involved in a model including spatio-temporal interaction, and for which the data are poorly informative due to the very low counts observed throughout the time-series. Note, however, that our analysis of the marginal spatial and temporal component processes makes no assumption about the dependence or otherwise between $S(x)$ and $U(t)$. Additionally, our objective was to estimate the regression coefficients which regulate the effects of purely spatial and temporal explanatory variables. For these reasons, we do not expect residual spatio-temporal interaction to affect our inferences.

Spatio-temporal mapping of plague incidence was not part of the objectives of the present analysis. However, the fitted model could be used to conduct predictive inference for the spatially continuous process $S(x)$ by first sampling from the joint predictive distribution of the district-wide averages S_i and then from the distribution of $S(x)$ at a set of prediction locations x, conditional on S_i. It would then be interesting to compare the accuracy of this approximate procedure with a different approximation used by Kelsall & Wakefield (2002) and with the exact method used by Diggle et al. (2013).

Authors contribution

EG conceived the study and wrote the first draft of the paper. EG and PD undertook all analyses. KK, MR, MR and MB provided access to the plague data, while CC provided climate data. All authors discussed the interpretation of the results, helped prepare subsequent drafts and approved the final version.

Acknowledgements

Emanuele Giorgi holds an MRC Biostatistics Fellowship (grant no. MR/M015297/1). Matthew Baylis acknowledges Leverhulme Trust Research Leadership Award (grant no. F/0025/AC). Cyril Caminade acknowledges support by The Farr Institute for Health Informatics Research (grant no. MR/M0501633/1).

A Computational details

In this section we give more details on the MCMC algorithms that were developed for the marginal temporal and spatial analyses of the data.

A.1 Priors

We use the following set of independent priors.

- $\log \alpha_1 \sim N(0, 10^3)$, $\log \alpha_2 \sim N(0, 10^3)$.
- $a_l \sim N(0, 10^3)$, $b_l \sim N(0, 10^3)$ for $l = 0, \ldots, 12$.
- $\beta_i \sim N(0, 10^3)$ for $i = 1, \ldots, 4$.
- $\log \nu^2 \sim N(0, 25)$.
- $\log \psi \sim N(1, 25)$.
- $\log r \sim N(0, 2.25)$.
- $\gamma \sim N(0, 10^3)$.
- $\log \sigma^2 \sim N(2.3, 1)$.
- $\log \phi \sim N(4.6, 1.2)$.

A.2 Temporal analysis

Use $\eta(t)$ to denote the linear predictor given by

$$\log \alpha_2 + \log b(t) + U(t).$$

Let $\eta^\top = (\eta(1), \ldots, \eta(T))$, $\xi^\top = (\log \alpha_2, a_0, \ldots, a_{12}, b_0, \ldots, b_{12}, \beta_1, \ldots, \beta_4)$ and, finally, use ω to denote all the positive model parameters on the log-scale. At the j-th iteration of the MCMC algorithm, we update each of the three blocks ω, ξ and η separately as follows.

1. Use independent random walk Metropolis-Hastings algorithms to update each of the element of ω, say ω_i. A new value $\omega_{i(j)}$, is proposed from a Gaussian distribution with mean $\omega_{i(j-1)}$ and variance $h^2_{i(j)}$ which is adaptively tuned as

$$h_{i(j)} = h_{i(j-1)} + c_1 j^{c_2}(\alpha_j - 0.45),$$

where c_1, c_2 are pre-defined constants, α_j is the acceptance rate at iteration j and 0.45 is the acceptance rate suggested by Gelman et al. (1996) under the assumption of a Gaussian posterior for ω_i.

2. Let V denote the inverse correlation matrix for the random effects η. Then, update ξ using a Gibbs sampling scheme. The full conditional distribution of ξ is Gaussian with covariance matrix

$$\Omega_\xi = (10^{-3} I + D^\top V D/\nu^2)^{-1},$$

where D is a matrix of covariates, and mean

$$\mu_\xi = \Omega_\xi D^\top V \eta/\nu^2.$$

3. Update η using a Hamiltonian Monte Carlo algorithm (Neal, 2011). Specifically, let $H(\eta, w)$ be the Hamiltonian function

$$H(\eta, w) = w^\top w/2 - \log \pi(\eta|\theta, y),$$

where $w \in \mathbb{R}^T$ is the momentum variables vector. The partial derivatives of $H(\eta, w)$ determines how η and w change over time j according the the Hamiltonian equations

$$\begin{aligned} \frac{d\eta(i)}{dj} &= \frac{\partial H}{\partial w_i} \\ \frac{dw_i}{dj} &= -\frac{\partial H}{\partial \eta(i)}, \end{aligned}$$

for $i = 1, \ldots, T$. In order to implement the Hamiltonian dynamic, the above differential equations are discretized using the *leapfrog method* (Neal, 2011, pp. 120-121) and approximate solutions are then found.

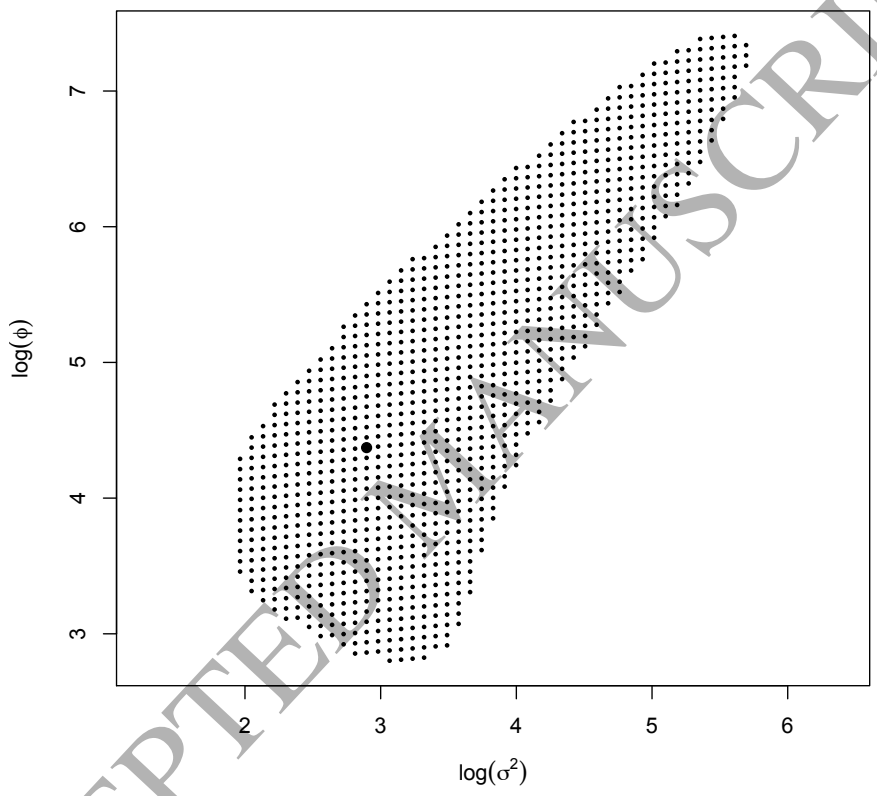

Figure 9: The set of points $\omega_{s,j} \in \mathcal{H}$ over which the posterior for ω_s is approximated; the larger point represents the mode $\hat{\omega}_s$.

A.3 Spatial analysis

Let $\omega_s^\top = (\log \sigma^2, \log \phi)$ and $W^\top = (\log \alpha_1, \gamma, S_1, \ldots, S_d)$. To obtain posterior samples for ω_s and W, we use the following MCMC algorithm which updates ω_s and W separately.

1. To update ω_s, we use a proposal distribution obtained through the INLA procedure (Rue, Martino & Chopin, 2009). Let y_s denote the vector of counts at each district; the proposal distribution for ω_s is then given by

$$\pi(\omega_s|y_s) \approx \pi_{\text{INLA}}(\omega_s|y_s) = \frac{\pi(\omega_s, \hat{W}(\omega_s)|y_s)}{\pi_G(\hat{W}(\omega_s)|\omega_s, y_s)},$$

   where $\pi_G(W|\omega_s, y_s)$ is the Gaussian approximation to the full conditional of W and $\hat{W}(\omega_s)$ is the mode of the full conditional for W, for a given ω_s. More details on how $\pi_G(W|\omega_s, y_s)$ is obtained can be found in Section 2.2 of Rue, Martino & Chopin (2009).

   After computing the mode of $\pi_{\text{INLA}}(\omega_s|y_s)$, say $\hat{\omega}_s$, we further approximate $\pi(\omega_s|y_s)$ by restricting its computation over a discrete set of points, shown in Figure 9, defined as

$$\mathcal{H} = \{\omega_{s,j} : \log\{\pi_{\text{INLA}}(\hat{\omega}_s|y_s)\} - \log\{\pi_{\text{INLA}}(\omega_{s,j}|y_s)\} < 10\}\,.$$

   At each iteration of the MCMC, we propose a new value for ω_s by simulating from the discrete bivariate distribution in Figure 9 where each point $\omega_{s,j}$ has probability mass

$$\frac{\pi_{\text{INLA}}(\omega_{s,j}|y_s)}{\sum_{\omega_{s,h}\in\mathcal{H}} \pi_{\text{INLA}}(\omega_{s,h}|y_s)}.$$

2. A new value for W is proposed using a random walk Metropolis-Hastings algorithm. The proposal distribution is given by a multivariate Gaussian with zero-mean and covariance matrix $h^2\Sigma(\omega_{s,j})$. The parameter h is tuned to obtain an acceptance rate of 0.234 using the same adaptive scheme in the first step of the previous section. Finally, $\Sigma(\omega_{s,j})$ corresponds to the inverse of the negative Hessian of the full-conditional of W, computed at $\hat{W}(\omega_{s,j})$. Note that the matrices $\Sigma(\omega_{s,j})$ are all pre-computed and stored in the previous step.

References

Amante, C. & Eakins, A. C. (2009). Etopo1 1 arc-minute global relief model: Procedures, data sources and analysis. Tech. Rep. NOAA Technical Memorandum NESDIS NGDC-24, National Geophysical Data Center.

Andrianaivoarimanana, V., Kreppel, K., Elissa, N., Duplantier, J., Carniel, E., Rajerison, M. & Jambou, R. (2013). Understanding the persistence of plague foci in Madagascar. *PLoS Neglected Tropical Disisease* **7**, e2382. DOI:10.1371/journal.pntd.0002382.

Ben Ari, T., Neerinckx, S., Gage, K. L., Kreppel, K., Laudisoit, A., Leirs, H. & Stenseth, N. C. (2011). Plague and climate: Scales matter. *PLoS Pathogens* **7**, e1002160. DOI:10.1371/journal.ppat.1002160.

Brygoo, E. R. (1966). Epidemiologie de la peste à Madagascar. *Archives de l'Institut Pasteur de Madagascar* **35**, 9–147.

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society, Series B* **17**, 129–164.

Diggle, P. J., Moraga, P., Rowlingson, B. & Taylor, B. M. (2013). Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm. *Statististical Science* **28**, 542–563.

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O. & Baptista, S. R. (2015). Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4. *Papers in Applied Geography* **1**, 226–234.

Gelman, A., Roberts, G. O. & Gilks, W. R. (1996). Efficient metropolis jumping rules. In *Bayesian Statistics*, J. M. Bernado, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds., vol. 5. Cambridge, MA: Oxford University Press, pp. 599–608.

Held, L., Hhle, M. & Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* **5**, 187–199.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R. & Joseph, D. (1996). The ncep/ncar 40-year reanalysis project. *Bullettin of the American Metereological Society* **77**, 437–471.

Kelsall, J. & Wakefield, J. (2002). Modeling spatial variation in disease risk: A geostatistical approach. *Journal of the American Statistical Association* **97**, 692–701.

Kreppel, K. S., Caminade, C., Telfer, S., Rajerison, M., Rahalison, L., Morse, A. & Baylis, M. (2014). A non-stationary relationship between global climate phenomena and human plague incidence in Madagascar. *PLoS Neglected Tropical Disisease* **8**, e3155. DOI:10.1371/journal.pntd.0003155.

Kreppel, K. S., Telfer, S., Rajerison, M., Morse, A. & Baylis, B. (2016). Effect of temperature and relative humidity on the development times and survival of Synopsyllus fonquerniei and Xenopsylla cheopis, the flea vectors of plague in Madagascar. *Parasites and Vectors* In press.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones & X.-L. Meng, eds., chap. 5. Chapman & Hall, CRC Press, pp. 113–162.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society, Series B* **71**, 319–392.

Stenseth, N. C., Samia, N. I., Viljugrein, H., Kausrud, K. L., Begon, M., Davis, S., Leirs, H., Dubyanskiy, V. M., Esper, J., Ageyev, V. S., Klassovskiy, N. L., Pole, S. B. & Chan, K.-S. (2006). Plague dynamics are driven by climate variation. *Proceedings of the National Academy of Sciences* **103**, 13110–13115.

Wakefield, J. & Lyons, H. (2010). Spatial aggregation and the ecological fallacy. In *Handbook of Spatial Statistics*, A. E. Gelfand, P. J. Diggle, P. Guttorp & M. Fuentes, eds. Chapman & Hall/CRC handbooks of modern statistical methods., pp. 541–558.

WHO (2009). Human plague: review of regional morbidity and mortality, 20042009. *Weekly Epidemiological Record* **85**, 40–45.